

MIRU-VNTR_{plus}: a web tool for polyphasic genotyping of *Mycobacterium tuberculosis* complex bacteria

Thomas Weniger^{1,*}, Justina Krawczyk², Philip Supply³, Stefan Niemann² and Dag Harmsen¹

¹Department of Periodontology, University Hospital Münster, Münster, ²Molecular Mycobacteriology, Forschungszentrum Borstel, Borstel, Germany and ³CIIL – Center for Infection and Immunity of Lille INSERM U 1019, CNRS UMR 8204 Univ Lille Nord de France Institut Pasteur de Lille, France

Received January 28, 2010; Revised April 15, 2010; Accepted April 22, 2010

ABSTRACT

Harmonized typing of bacteria and easy identification of locally or internationally circulating clones are essential for epidemiological surveillance and disease control. For *Mycobacterium tuberculosis* complex (MTBC) species, multi-locus variable number tandem repeat analysis (MLVA) targeting mycobacterial interspersed repetitive units (MIRU) has been internationally adopted as the new standard, portable, reproducible and discriminatory typing method. However, no specialized bioinformatics web tools are available for analysing MLVA data in combination with other, complementary typing data. Therefore, we have developed the web application MIRU-VNTR_{plus} (<http://www.miru-vntrplus.org>). This freely accessible service allows users to analyse genotyping data of their strains alone or in comparison with a reference database of strains representing the major MTBC lineages. Analysis and comparisons of genotypes can be based on MLVA-, spoligotype-, large sequence polymorphism and single nucleotide polymorphism data, or on a weighted combination of these markers. Tools for data exploration include search for similar strains, creation of phylogenetic and minimum spanning trees and mapping of geographic information. To facilitate scientific communication, an expanding genotype nomenclature (MLVA MtbC15-9 type) that can be queried via a web- or a SOAP-interface has been implemented. An extensive documentation guides users through all application functions.

INTRODUCTION

The bacteria of the *Mycobacterium tuberculosis* complex (MTBC) are the causative agents of tuberculosis (TB). This infectious disease is responsible for approximately 2 million deaths annually and foci of multi- and extensive drug resistance are emerging worldwide (1). Genotyping of MTBC strains empowers epidemiological surveillance and control, e.g. by permitting detection of clonal spread of multi-drug resistant strains and unsuspected outbreaks. At the clinical level, molecular typing enables identification of false positive cases due to laboratory cross-contamination and distinction between exogenous reinfection and relapse from initial infection. From a research perspective, molecular strain typing is valuable for the deciphering of the MTBC population structure and thus enabling the development of new diagnostics, vaccines or treatments (2).

The previous gold standard for MTBC genotyping—IS6110 restriction fragment length polymorphism (RFLP) typing—is laborious and time consuming and the resulting complex banding patterns make inter-laboratory comparisons difficult (3). As the MTBC gene sequence variation is low classical sequence-based typing methods such as multi-locus sequence typing, even extended to tens of genes, are only informative for the identification at genetic (sub-)lineage but not at strain level (4,5). Likewise, large sequence polymorphisms (LSP) or single nucleotide polymorphisms (SNP) elucidated MTBC genetic lineages differing in their geographical distribution immunogenicity, virulence and association with multidrug resistant TB (6–8).

However, the use of other, more variable markers is needed for a desirable finer phylogenetic classification and detection down to strain level to the advantage of public health and clinical objectives. As a portable,

*To whom correspondence should be addressed. Tel: +49 251 83 49882; Fax: +49 251 83 47134; Email: tweniger@uni-muenster.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2010. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.5>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

reproducible and discriminatory typing tool, the multi-locus variable number tandem repeat (VNTR) analysis (MLVA) targeting a number of tandem repeat loci including the mycobacterial interspersed repetitive units (MIRU) has, therefore, been internationally adopted as the new standard (9,10). Spoligotyping is often used as a rapid additional typing method to increase the discriminatory power of MIRU-VNTR typing. This method addresses the presence or absence of 43 spacers at the so-called direct repeat locus, a paradigmatic member of the family of clustered regularly interspaced short palindromic repeats (CRISPR) (11,12). In addition to their combined discriminatory power, 24-locus-based MIRU-VNTR typing and spoligotyping data are predictive of the main MTBC phylogenetic lineages, although they are somewhat less deterministic for such purposes than are LSP or sequence-based data (13,14).

Several web services containing strain collections of MLVA or spoligotyping data are available, e.g. MLVAbank or spolTools and SITVIT/SpolDB4 (15–18). Each of these web applications offers functions for simple data comparison and basic clustering for only one genotyping method. However, there was no online tool available that provided a comprehensive analysis of MTBC strains based on multiple genotyping methods and tools for directly naming new genotypes. We described recently such a system, i.e. MIRU-VNTR_{plus}, with a focus on evaluation of a reference database that allows a robust-lineage identification based on the combination of different genotyping data (19). Here, we describe for the first time the complete functionality of MIRU-VNTR_{plus}, including the new features minimum spanning tree (MST), geographic mapping and the nomenclature service.

IMPLEMENTATION

The MIRU-VNTR_{plus} server was implemented in Java using the JavaServer Faces technology together with the RichFaces AJAX extension for creating the web pages. Java Servlets were used for image rendering. In order to store the data of the provided nomenclature service persistently, a MySQL database was used. Database access was implemented by means of the object-relational mapping framework Hibernate, while the SOAP interface for accessing the nomenclature service was realized by means of the Apache Axis2 framework. The mapping function uses Google Maps for geocoding and display of an interactive map. MIRU-VNTR_{plus} is compatible with the most common web browsers, i.e. Mozilla Firefox 2 and 3; Microsoft Internet Explorer 6, 7 and 8; Opera 9 and 10; Apple Safari 4; and Google Chrome 3. To enable the application's full functional capability JavaScript has to be activated.

The application allows the input of copy numbers for 24 MIRU-VNTR loci and 15 or 12 loci subsets (10). Furthermore, the following typing data are supported: presence or absence of 43 spoligotyping spacers, presence or absence of 15 standard LSPs (6,7), values of

six SNPs (20) and susceptibility data for 16 drugs. Finally, in addition to general descriptive information (e.g. strain ID and country of isolation) input for three user-specific data fields can be supplied. Furthermore, a reference database containing genotyping data of 186 strains representing the major MTBC lineages is available (19). IS6110 RFLP fingerprinting images are also shown for visual comparison of these strains.

Distance measures that can be used for strain comparison include categorical distance, D_C , $(\delta\mu)^2$ and D_{SW} (21–23). Double and variant alleles are regarded as own categories for categorical distance calculation and are treated as missing data for all other distance measures. When used for pairwise strain comparisons, the categorical distance is identical to the distance measure D_A that is well suited for phylogenetic analysis of VNTR markers (24). Polyphasic typing is achieved by calculating combined pairwise distances for all chosen typing methods. A combined distance is calculated by multiplying the distance of each method by a weighting factor, summing up the distances for all methods, and dividing this value by the sum of the weighting factors. Missing data is ignored for the pairwise distance calculation, thus allowing users to work with a self-defined subset of loci.

Phylogenetic trees are inferred using the un-weighted pair group method with arithmetic means (UPGMA) or neighbour-joining (NJ) algorithms (25,26). Generated trees can be re-rooted using a manually selected outgroup. The resulting trees can be exported to Newick and NEXUS format and the underlying distance matrix can be downloaded as a MEGA file. A minimum spanning tree is created using Kruskal's algorithm and a force-directed graph layout for visualization (27). All trees can be downloaded as raster image (PNG), vector image (SVG, EMF) or PDF document.

RESULTS

The freely accessible MIRU-VNTR_{plus} web application (<http://www.miru-vntrplus.org>) offers three main functions: (i) phylogenetic lineage identification by using a reference database; (ii) analysis and visualization of genotyping data; and (iii) access to the MLVA MtbC15-9 nomenclature service (Figure 1). An extensive documentation including a manual, multi-media tutorials and protocols for the genotyping methods complements the web server.

Data input

Genotyping data can be entered for a single strain via a web form, copied from a spreadsheet application via clipboard or uploaded by using Comma Separated Values (CSV) or Microsoft (MS) Excel files. Template CSV and MS Excel files that simplify the upload process are available. MIRU-VNTR_{plus} allows users to upload data for up to 500 strains. Various formats for genotyping data are accepted, e.g. MIRU-VNTR alleles with incomplete tandem repeat units (variant alleles) and two alleles simultaneously detected for a given locus (double alleles),

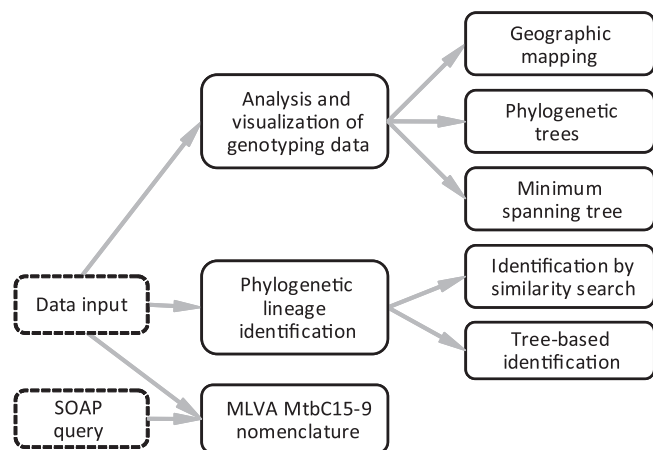


Figure 1. The three main functions of MIRU-VNTR_{plus} are (i) phylogenetic lineage identification by using a reference database; (ii) analysis and visualization of genotyping data; and (iii) access to the MLVA MtbC15-9 nomenclature service.

CDC notation for MIRU-VNTR data (by using 'A' for 10 repeats, 'B' for 11 repeats, etc. double digits are avoided) or binary and octal numbers for spoligotyping patterns. After exclusion of possible PCR artefacts (i.e. classical stutter peaks) concordantly observed double alleles in several independent VNTR loci for a given sample indicate the presence of a mixed DNA population. This mixed population can result from a true mixed infection, or from culture or DNA contamination. In contrast, the occurrence of double alleles in a single locus rather suggests the presence of a given allelic variant within a clonal isolate. In the following, the entered or uploaded strain data is referred to as 'user strains'.

Analysis and visualization of genotyping data

MIRU-VNTR_{plus} supports polyphasic typing by calculating a combined distance for different types of genotyping data. An input form allows the selection of the used genotyping methods, distance measures and weightings (Figure 2A). This form can be accessed on all pages that use calculated distances. A data table that can be sorted by any column displays the strain data. Using the UPGMA or NJ algorithm a dendrogram can be directly included in this table, thus ordering the strains by their position in the tree (Figure 2B). A filter function allows one to exclude strains from analysis that do not match certain criteria. Complex filtering criteria can be created by combining comparison operators for certain data fields with the logical operators AND or OR. Strains can be marked manually or automatically with a background colour, e.g. according to the value of lineage, genetic marker or user data. By selecting a strain as the comparison strain all genetic marker differences are highlighted for all other strains. In addition the distances between the comparison strain and all other strains are displayed for sorting and filtering.

Phylogenetic trees can be created and modified interactively on an extra page. Depending on the user selection, trees are displayed as dendrograms or radial graphs.

The label text can be chosen and genotyping data can be added to the image. It is possible to export the distance matrix, the tree and the resulting image to various formats. Clicking a branch or leaf of the tree opens a pop-up window that shows the genotyping data for all sub-tree strains and offers to re-root the tree, swap branches and mark sub-tree strains by colour as interactive features. In addition to classical phylogenetic trees, MIRU-VNTR_{plus} also allows the calculation of MSTs. However, MSTs can be drawn based on data from only one typing method at a time. Clonal complexes (CC) as defined by genotypes sharing a selectable maximum locus difference are highlighted in the MST image. The choice of the label text, length of connection lines and zoom factor modifies the appearance of the MST. Using the context menu strains can again be marked by colour.

The strain-mapping feature displays the geographical distribution of strains on a map. The location information is retrieved from the data field country of isolation that might as well include ZIP codes or city information. The distribution of species, lineages or user data field values for each location is visualized in a colour-coded pie chart.

Phylogenetic lineage identification

The comparison with the reference database allows the identification of the phylogenetic lineage of user strains. Again, the identification can be based on a combined distance of several genotyping methods. Identification by similarity search lists reference database strains displaying an adjustable maximum distance to each user strain. The default maximum distance has been determined on the basis of validation tests using an external strain data set (19). The similarity search results can be used for the automatic assignment of species and lineage information to user strains. All results are exportable as MS Excel or CSV file. In most cases, identification by similarity search will not be sufficient to undoubtedly identify all user strains. Here, an additional tree-based analysis can be carried out by calculating a UPGMA or NJ tree that contains all user and reference database strains. Investigating strains that are monophyletic (i.e. ingroups) compared to reference database strains in the different tree branches allow the user to infer species and lineage assignment. Such assignments may then be set by clicking strains or nodes on the tree (Figure 3).

MtbC15-9 nomenclature service

For the exchange of MIRU-VNTR data, the reporting of the full genotype with copy numbers of each locus in perfect order is mandatory. To facilitate scientific communication, MIRU-VNTR_{plus} introduces an expanding nomenclature that assigns a numerical code to MIRU-VNTR patterns. The MLVA MtbC15-9 type is a juxtaposition of two subtypes, i.e. the MtbC15 and MtbC9 type. These types are based on a set of the 15 most discriminatory MIRU-VNTR loci and a set of nine ancillary loci, as inferred by the analysis of single locus variation frequencies on a large international strain collection (10). The web application provides forms for converting

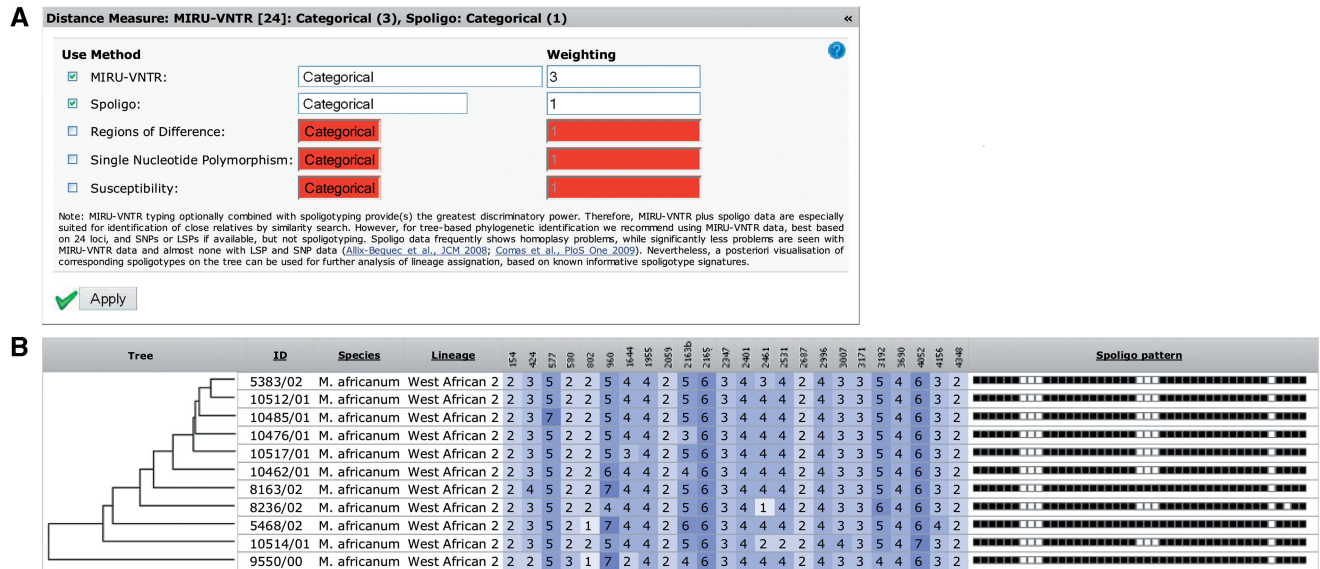


Figure 2. (A) Input form allowing the selection of the used genotyping methods, distance measures and weightings. **(B)** Data exploration table. The data for each strain is displayed in a single row. Next to the ID, species and lineage columns, the table contains MIRU-VNTR data highlighted in blue shades, indicating the different copy numbers and spoligotyping patterns. An UPGMA tree is used for ordering the strains.

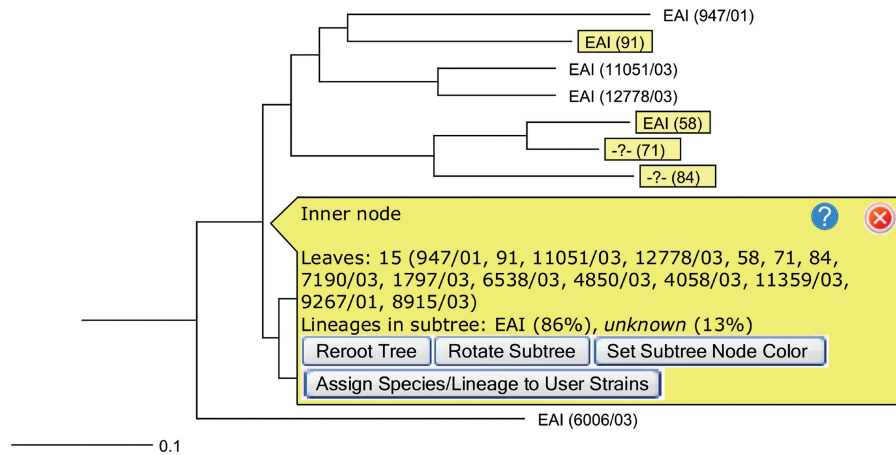


Figure 3. Tree-based identification. Clicking a branch or leaf of the tree opens a popup window that offers functions for re-rooting the tree, swapping of branches and marking subtree strains by colour. Furthermore, species/lineage may be assigned for user strains (labelled with yellow boxes).

MtbC15-9 types into VNTR copy numbers and vice versa either in single or batch mode. The results of these queries can be exported as CSV or MS Excel files. When uploading user strains MIRU-VNTRplus automatically detects and displays known MtbC15-9 types. New MtbC15-9 types can be added to the nomenclature service provided that contact information is supplied. An email that contains a link for confirmation of the assigned types is sent to the given contact email address. On the condition of confirmation, the new MtbC15-9 types are added to the nomenclature server and made publicly available. External applications may query the nomenclature service via a SOAP interface. The loci used for the MtbC15 and MtbC9 nomenclature, all existing types, the copy numbers for specific MtbC15 or MtbC9 types and the types for a given MIRU-VNTR pattern can be retrieved by SOAP queries. A documentation of the SOAP

service that includes five examples in three different programming languages (Java, Perl and C#) is available for developers.

APPLICATION EXAMPLE

To provide an application case, a recently published data set containing MIRU-VNTR data and spoligotyping patterns for 97 MTBC strains has been re-analysed (28). These data can be loaded from the homepage of MIRU-VNTRplus by clicking the 'open example dataset' link. To reproduce the published results only the MIRU-VNTR typing data for all 24 loci and categorical distance measure must be chosen for distance calculation. A phylogenetic lineage identification is obtained for 39 of the 97 strains when a similarity search is

carried out by using the menu command 'Automatically Set Best Matching Species and Lineage' with a distance threshold of 0.17 (default value). By means of the tree-based identification, lineages of additional user strains can be derived. By drawing a NJ tree that is re-rooted at the branch containing the two *M. canetti* reference database strains (outgroup), species and lineage information can be reliably determined for 17 additional strains (one strain with lineage West African 1, three strains with lineage S, one strain with Beijing lineage, eight strains with lineage LAM and four strains with lineage EAI). A closer look at the remaining unidentified strains clearly reveals the presence of the two lineages (strains with ID 19, 25, 35, 40, 52, 42, 63 and ID 33, 78, 83, 56, 34, 72, 50, 51, 12, 79), which had been described as 'Sierra Leone-1' and 'Sierra Leone-2' in the original publication. The creation of a MST using the 24 MIRU-VNTR loci and a maximum locus difference of four within a CC reproduces the CCs of the publication, however, with a slightly different visualization and ordering. Remarkably, MST groups the 'Sierra Leone-1' and 'Sierra Leone-2' strains into specific CCs. The key findings of Homolka *et al.* (28) are thus confirmed by using MIRU-VNTR_{plus}. The data exploration table gives an overview of all strains and allows further data exploration. Possible new MtbC15-9 types can be submitted by choosing the menu command 'Assign MLVA MtbC15-9 Types'.

CONCLUSION

The MIRU-VNTR_{plus} application allows quick identification of lineages for MTBC strains and exploration of data based on a combination of up to four different genotyping methods. Furthermore, an expanding nomenclature service for MLVA MtbC types has been established. It is planned to extend the reference database by adding further quality-controlled data sets. Future developments will include an open, extendable database with data sets from other researchers that can be used in addition to the reference database. Since MLVA, CRISPR and SNP typing schemes are being published for an increasing number of especially monomorphic pathogenic bacterial species, MIRU-VNTR_{plus} may serve as a model for a powerful, generalized tool to analyse genotypes using these markers and other categorical data.

FUNDING

German Federal Ministry of Education and Research in the framework of the Network Zoonoses (grant number 01KI07124 to D.H. for development of the SOAP-interface); PathoGenomikPlus Network (grant number 0313801J to S.N. for collecting strains and genotyping). Funding for open access charge: German Federal Ministry of Education and Research (grant number 01KI07124 to D.H.).

Conflict of interest statement: P. Supply has declared a potential conflict of interest. P. Supply is a consultant

for Genoscreen, Lille, France. All other authors have declared that no competing interests exist.

REFERENCES

1. World Health Organization. (2009) Global tuberculosis control. *WHO Report*. World Health Organization, Geneva, Switzerland.
2. Gagneux, S. and Small, P.M. (2007) Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect. Dis.*, **7**, 328–337.
3. van Embden, J.D., Cave, M.D., Crawford, J.T., Dale, J.W., Eisenach, K.D., Gicquel, B., Hermans, P., Martin, C., McAdam, R. and Shinnick, T.M. (1993) Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J. Clin. Microbiol.*, **31**, 406–409.
4. Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S. and Musser, J.M. (1997) Restricted structural gene polymorphism in the *Mycobacterium tuberculosis* complex indicates evolutionarily recent global dissemination. *Proc. Natl Acad. Sci. USA*, **94**, 9869–9874.
5. Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W. *et al.* (2008) High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.*, **6**, e311.
6. Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K. *et al.* (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl Acad. Sci. USA*, **99**, 3684–3689.
7. Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C. *et al.* (2006) Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA*, **103**, 2869–2873.
8. Hirsh, A.E., Tsolaki, A.G., DeRiemer, K., Feldman, M.W. and Small, P.M. (2004) Stable association between strains of *Mycobacterium tuberculosis* and their human host populations. *Proc. Natl Acad. Sci. USA*, **101**, 4871–4876.
9. Supply, P., Magdalena, J., Himpens, S. and Locht, C. (1997) Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol. Microbiol.*, **26**, 991–1003.
10. Supply, P., Allix, C., Lesjean, S., Cardoso-Oelemann, M., Rüsch-Gerdes, S., Willery, E., Savine, E., de Haas, P., van Deutekom, H., Roring, S. *et al.* (2006) Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.*, **44**, 4498–4510.
11. Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M. *et al.* (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.*, **35**, 907–914.
12. Oelemann, M.C., Diel, R., Vatin, V., Haas, W., Rüsch-Gerdes, S., Locht, C., Niemann, S. and Supply, P. (2007) Assessment of an optimized mycobacterial interspersed repetitive-unit-variable-number tandem-repeat typing system combined with spoligotyping for population-based molecular epidemiology studies of tuberculosis. *J. Clin. Microbiol.*, **45**, 691–697.
13. Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S. *et al.* (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathog.*, **4**, e1000160.
14. Comas, I., Homolka, S., Niemann, S. and Gagneux, S. (2009) Genotyping of genetically monomorphic bacteria: DNA sequencing in *Mycobacterium tuberculosis* highlights the limitations of current methodologies. *PLoS ONE*, **4**, e7815.
15. Le Flèche, P., Fabre, M., Denoeud, F., Koeck, J. and Vergnaud, G. (2002) High resolution, on-line identification of strains from the *Mycobacterium tuberculosis* complex based on tandem repeat typing. *BMC Microbiol.*, **2**, 37.

16. Grissa, I., Bouchon, P., Pourcel, C. and Vergnaud, G. (2008) On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. *Biochimie*, **90**, 660–668.
17. Tang, C., Reyes, J.F., Luciani, F., Francis, A.R. and Tanaka, M.M. (2008) spolTools: online utilities for analyzing spoligotypes of the *Mycobacterium tuberculosis* complex. *Bioinformatics*, **24**, 2414–2415.
18. Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajj, S.A., Allix, C., Aristimuño, L., Arora, J., Baumanis, V. *et al.* (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. *BMC Microbiol.*, **6**, 23.
19. Allix-Béguec, C., Harmsen, D., Weniger, T., Supply, P. and Niemann, S. (2008) Evaluation and strategy for use of MIRU-VNTR_{plus}, a multifunctional database for online analysis of genotyping data and phylogenetic identification of *Mycobacterium tuberculosis* complex isolates. *J. Clin. Microbiol.*, **46**, 2692–2699.
20. Niemann, S., Harmsen, D., Rüsche-Gerdes, S. and Richter, E. (2000) Differentiation of clinical *Mycobacterium tuberculosis* complex isolates by *gyrB* DNA sequence polymorphism analysis. *J. Clin. Microbiol.*, **38**, 3231–3234.
21. Cavalli-Sforza, L.L. and Edwards, A.W. (1967) Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.*, **19**, 233–257.
22. Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L. and Feldman, M.W. (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl Acad. Sci. USA*, **92**, 6723–6727.
23. Shriver, M.D., Jin, L., Boerwinkle, E., Deka, R., Ferrell, R.E. and Chakraborty, R. (1995) A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.*, **12**, 914–920.
24. Takezaki, N. and Nei, M. (2008) Empirical tests of the reliability of phylogenetic trees constructed with microsatellite DNA. *Genetics*, **178**, 385–392.
25. Sokal, R.R. and Michener, C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **38**, 1409–1438.
26. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
27. Kruskal, J.B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.*, **7**, 48–50.
28. Homolka, S., Post, E., Oberhauser, B., George, A.G., Westman, L., Dafaie, F., Rüsche-Gerdes, S. and Niemann, S. (2008) High genetic diversity among *Mycobacterium tuberculosis* complex strains from Sierra Leone. *BMC Microbiol.*, **8**, 103.